CASG：A Conditional Autoregressive Co-Speech Gesture Generation Network for Semantics

Wangyang Tuo

***Abstract*—Co-speech gesture generation, a subset of 3D motion generation, aims to generate appropriate motion sequences from audio or other conditions. While many existing methods focus on the rhythm between motion and audio, they often neglect the semantics of gestures. Furthermore, approaches based on diffusion models or Transformer require significant time for training and inference, making them unsuitable for real-time applications. In this paper, we proposed CASG, a network based on the conditional autoregressive model, which effectively enhances the semantics of generated gestures through a semantic enhancement module inspired by VQ-VAE. Additionally, the loss function is improved for 3D rotational and translational transformations of motion sequences, addressing the instability issue in generated models. Extensive experiments demonstrate that our model outperforms competing methods in terms of semantics, rhythm and stability.**

***Index Terms*—gesture genaration, semantics, VQ-VAE, autoregressive model.**

I. Introduction

3D motion generation involves creating action sequences in three-dimensional space using computer algorithms and technologies. This process typically encompasses capturing, simulating, editing, and rendering action data to achieve realistic dynamic performance in various media such as movies, animations, video games, and virtual reality (VR).

Co-speech gesture generation, a specific branch of 3D motion generation, focuses on generating adaptive motion from a segment of audio. Early approaches relied on rule-based methods, utilizing predefined correspondences between human conversations, speech, and other states. However, these methods have limitations in terms of rule complexity, predefined action size, and the need for extensive manual work.

Co-speech gesture generation, a specific branch of 3D motion generation, focuses on generating adaptive motion from a segment of audio. Early approaches relied on rule-based methods, utilizing predefined correspondences between human conversations, speech, and other states. However, these methods have limitations in terms of rule complexity, predefined action size, and the need for extensive manual work.

In recent years, data-driven gesture generation methods have gained prominence. These methods require less manual effort and offer greater flexibility by dynamically generating new gestures and being easily scalable to large datasets. Deep learning techniques, including recurrent neural networks (RNN), long short-term memory networks (LSTM), generative adversarial networks (GAN), and diffusion models, have been extensively used for gesture generation from speech. CaMN [1] designed a cascade structure to drive the generation of poses based on facial, body, audio, text transcript and speaker id. ZeroEGGS [2] adds stylization to the dataset, and uses the variational framework to learn gesture embedding, so that gesture can be modified through potential spatial operations or mixing and scaling of style embedding. DiffuseStyleGesture [3] introduces cross local attention and self-attention into the diffusion model to generate better audio-matching and real gestures. However, simply using text as input can not deeply understand the semantic information. Therefore, some methods adopt specific structures to better learn the semantic information of audio, so as to achieve the gesture expression effect with more semantic information. LivelySpeaker [4] uses the text as a semantic description based on the diffusion model. The motion sequence is divided into fixed segments, and an encoder-decoder structure is input to calculate the reconstruction loss, while the loss calculation is compared with the text vector. GestureDiffuCLIP [5] introduces the CLIP structure into the diffusion model to learn the mapping relationship between text and motion sequences in the latent space, so that the generated gestures can realize the semantic information.

Many existing methods primarily utilize single audio inputs, potentially neglecting the semantic information present in the audio. Some approaches incorporate text transcriptions as part of the input, while others employ specific structures to better learn speech semantics, resulting in gesture expressions with richer semantic information. However, these methods often suffer from slow training and inference times, limiting their applicability in real-time scenarios. To address these limitations, we propose a semantic-enhanced co-speech gesture generation method. Our approach utilizes a semantic enhancement module SEM based on VQ-VAE to establish a mapping relationship between text vectors extracted by CLIP and vectors in the VQ-VAE latent space, thereby enhancing the model’s semantic understanding. An improved structure optimizes the extraction and fusion of multimodal information, ensuring both semantic richness and efficient reasoning. Additionally, we optimize the loss function to improve the stability of the generated gestures.

In summary, our contributions are as follows:

1) We propose a semantic enhancement module SEM, which establishes a semantic mapping relationship between

Manuscript received March 03, 2025

Wangyang Tuo, School of software, Tiangong University, Tianjin,China

action sequences and text transcripts, enhancing the understanding of semantic information.

2) We propose CASG, a fast pose generation model, which can more accurately extract and fuse multimodal information. Moreover, we enhance the loss function for the rotation and translation of motion capture data, which aids in generating more diverse, natural, and stable high-quality gestures.

3) Experiments conducted on the BEAT dataset have shown that our model outperforms competing methods in terms of semantic representation, naturalness, and stability. These results underscore the effectiveness of our approach in generating realistic and meaningful co-speech gestures.

II. Related Work

Human motion synthesis. Human motion synthesis has a rich history, while human motion prediction being one of the most captivating fields. This domain aims to predict future motion based on past motion sequences. Existing methods incorporate spatial and temporal information to generate future motion sequences. Deep neural networks, with their formidable modeling capabilities, have been extensively employed in gesture generation. Traditional models, such as MLP, RNN and Transformer, have been utilized alongside generative models like VAE, diffusion models, or flow-based models. Co-speech generation is a sub-task within human motion generation, focusing on generating 3D human motion in response to various conditions. MotionCLIP [6] leverages aligned text and motion embeddings, with a CLIP text encoder, and rendered images providing additional supervision. For basic motion generation, various methods have been proposed, including predefined motion classes as in GesGPT [7], or additional text encoders as in FreeTalker [8], and temporal motion combinations derived from a series of natural actions. However, these methods typically concentrate solely on rhythm. In contrast, our approach considers both the semantics and rhythm of gesture generation within a unified framework.

VQ-VAE. Vector Quantized-Variational Auto-encoder [9], is a generative model that combines the concepts of Variational Autoencoders (VAE) and vector quantization. Initially developed for image generation tasks, it features an autoencoder architecture designed to learn and reconstruct data using discrete representations. VQ-VAE begins by constructing a codebook, essentially an embedding space of features. The encoder processes the input image to extract its feature map. Each feature vector then finds the closest vector in the codebook, using the index to retrieve the closest vector and create a quantized feature map. This quantized map is passed to the decoder, which reconstructs the original image. Compared to VAE that sample and generate from a Gaussian distribution, VQ-VAE utilizes a limited codebook, making it easier for the decoder to handle the feature distribution in the hidden state while also constraining the variance. Sampling from a codebook is simpler than completely free sampling, making it more efficient for generating samples.

Due to its strong capabilities in data representation, generation and compression, VQ-VAE has been applied to various modalities beyond images. These include audio generation, style transfer, text motion generation, and co-speech gesture generation. For instance, T2M-GPT [10] employs a generation framework based on VQ-VAE and GPT, learning to generate human motion from high-quality discrete representations and enhancing the consistency between text and generated motion. MotionGPT [11] treats motion as a language, leveraging large-scale motion models and integrating language data to train a motion-related codebook. This codebook uses discrete quantization for human motion and converts 3D motion into motion tokens, allowing for a unified approach to modeling motion and text in language and learning their correlation. Another example is TM2D [12], which generates 3D motion from music and text. It proposes a cross-model transformer and a bimodal feature fusion strategy to encode audio and text features, utilizing the VQ-VAE framework to encode the motion of all training sets into a shared feature space.

III. Method

Our method consists of two main steps, aimed at enhancing the semantic expression of gesture generation. First, we train the semantic enhancement module SEM. This step involves training a semantic learning module based on VQ-VAE. The original action training data is encoded into a latent space. We then access the frozen CLIP [13] text encoder to obtain the vector representation of the transcript and calculate the cosine similarity loss between the vectors in the CLIP latent space z_{CLIP} and the action vectors in the VQ-VAE latent space z_{emb}. By learning to reconstruct the original action sequence from this potential space, we establish a semantic mapping relationship between the actions and the transcripts. Second, we train the co-speech gesture generate network. We extract features from both the audio and gesture data, concatenate them, and input them into the gesture generator. The audio features are divided into low-level and high-level components. The low-level part focuses on capturing audio rhythm, tempo, and other rhythmic information, represented by the Mel spectrogram extracted using librosa. The high-level part aims to extract deeper semantic information from the audio, utilizing a pre-trained WavLM [14] structure. During gesture generation, we leverage the VQ decoder trained in the first step to promote the semantic relevance and stability of the generated actions.

Semantic Enhancement Module. Considering the powerful role of VQ-VAE in image generation tasks, this paper utilizes a semantic enhancement module SEM based on the VQ-VAE structure. The module encodes the motion sequence into the latent space, learns to reconstruct the original motion sequence from a set of discrete representations, and calculates the cosine similarity with the frozen CLIP text encoder to establish the mapping relationship between the codebook and the text transcript.

As shown in Fig.1(a), the encoder and decoder of the semantic enhancement module are implemented as GRU structures. The encoder encodes the motion sequence into the latent space, while the decoder reconstructs the original motion sequence from a set of discrete codes. The codebook serves as an embedding layer, clustered into a predefined codebook size. This process discretizes the continuous vector

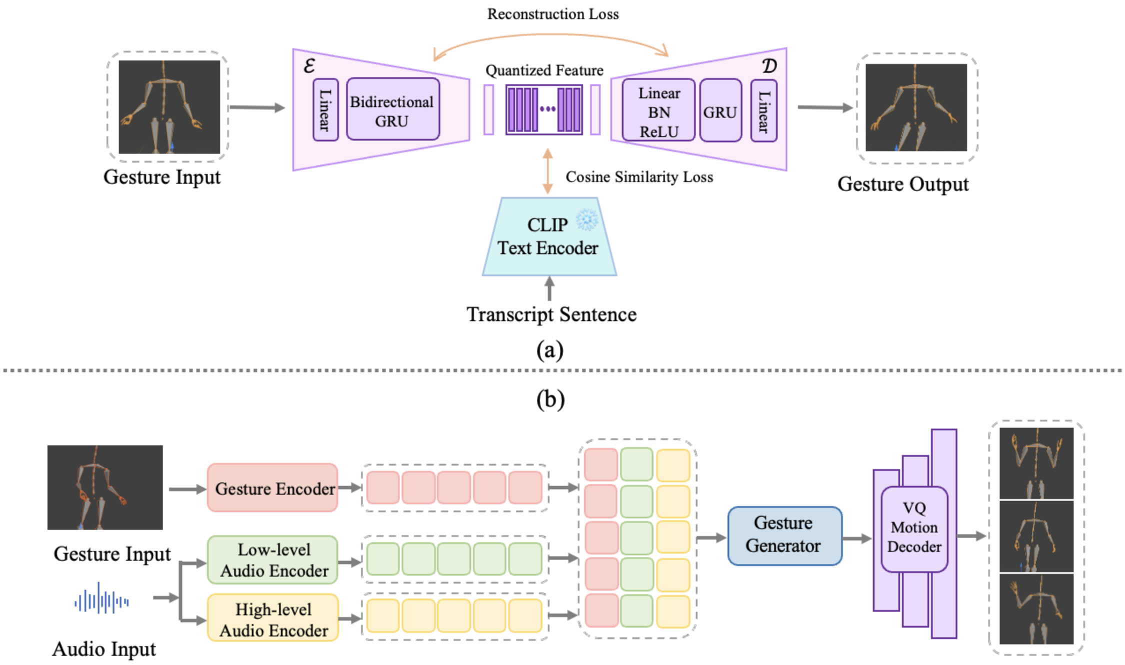

Fig. 1: Overview of the architecture. Our method is divided into two steps. First, as shown in (a), we train a semantic enhancement module based on VQ-VAE. This module learns to produce a semantically relevant VQ decoder by optimizing the loss function through the cosine similarity between the vectors in the latent space and the text vectors extracted by CLIP. The second step, as shown in (b), involves training a co-speech gesture generation network using a conditional autoregressive model, where the audio is segmented into two channels, each channel concentrating on different aspects of the audio: rhythm and semantics. When generating motion, the network utilizes the VQ decoder trained in the first stage, allowing the model's final prediction results to more effectively convey the semantics.

into a predefined number of latent vectors z_{emb}. During the decoding phase, the quantized vector $z_{emb}$is reconstructed into an action sequence through the GRU structure. The loss function for the reconstruction part is defined as follows:

$$L_{VQ} = \log p(x|z_q(x)) + ||sg[z_e(x)] - e||_2^2 + \beta||z_e(x) - sg[e]||_2^2 \quad (1)$$

The first term in the loss function is solely used for training the encoder and decoder, with identical gradients, and represents the reconstruction loss. The second term is dedicated to training the codebook to be closer to the embeddings Z_e. The third term trains the encoder while fixing the codebook gradient to ensure the encoder's output stability. Here, $sg[\cdot]$ denotes the gradient stop operation, x represents the input, $z_e(x)$ denotes the encoder vector of input x, and $z_q(x)$ represents the generated quantization vector, which serves as the input to the decoder.

Simultaneously, we employ cosine similarity loss L_{cos} during training to compute the cosine similarity between the CLIP semantic embedding z_{CLIP} and the latent space embedding z_{emb} , establishing the semantic relationship between text transcript and gestures:

$$L_{sem} = L_{cos}(z_{CLIP}, z_{emb}) \quad (2)$$

The complete training objectives of semantic enhancement
module are:

$$L_{full} = L_{VQ} + L_{sem} \quad (3)$$

Dual Audio Encoder. The audio encoder is divided into two parts, focusing on extracting both low-level and high-level audio information. The low-level audio encoder utilizes librosa to extract the Mel spectrogram from the original audio
data. Mel spectrograms offer an improved representation of frequency domain signals compared to traditional spectrograms. They employ a Mel frequency scale instead of a linear scale, making them more sensitive to low frequencies and aligning more closely with human auditory perception. Consequently, Mel spectrograms are widely used for audio information extraction. After obtaining the Mel spectrograms, the encoder employs a series of 1D convolutional layers, activation functions, and frame-wise linear layers to obtain the embedded vector sequence representing the low-dimensional audio information.

To extract high-level audio features, a pre-trained WavLM structure is employed. Upon inputting the original audio data, the high-level audio encoder first processes it through a structure consisting of multiple convolutional layers. Subsequently, the features extracted from these convolutional layers are passed into a time context module based on the Transformer architecture to capture higher-level voice information. The WavLM model transforms the original audio into a series of discrete tokens, which encapsulate rich high-level semantic information.

Gesture Encoder. The gesture encoder transforms the input motion capture data into a fixed-length embedded vector. The features extracted include the character's position, rotation, and speed relative to the local body transformation. All trajectories and body joint transformations are calculated relative to the root trajectory's transformation. To facilitate training, the original joint rotation is converted into a 2-axis rotation matrix representation, with the dimension changed to 6 * j, where j is the number of joints. Unlike quaternion or Euler angle representations, the rotation matrix uses relative forward and upward vectors to represent joint rotation, which is continuous, thereby avoiding quaternion interpolation issues during neural network training. Additionally, the normalization of these features further aids in neural network training. These features are calculated from the original Euler angles of the dynamic capture data. During reasoning, the output is converted back into Euler angles to restore the final motion. The joint and root rotation velocity are specified using the scale angle axis reference from [15]. Thus, each frame is composed of an vector $a = [\rho_p, \rho_r, v_p, v_r, \epsilon_p, \epsilon_r]$, where ρ_p and ρ_r are the translation and rotation positions of the joint, v_p and v_r are the velocity of the local translation and rotation of the joint, and ϵ_p and $\epsilon_p$are the velocity of the root translation and rotation.

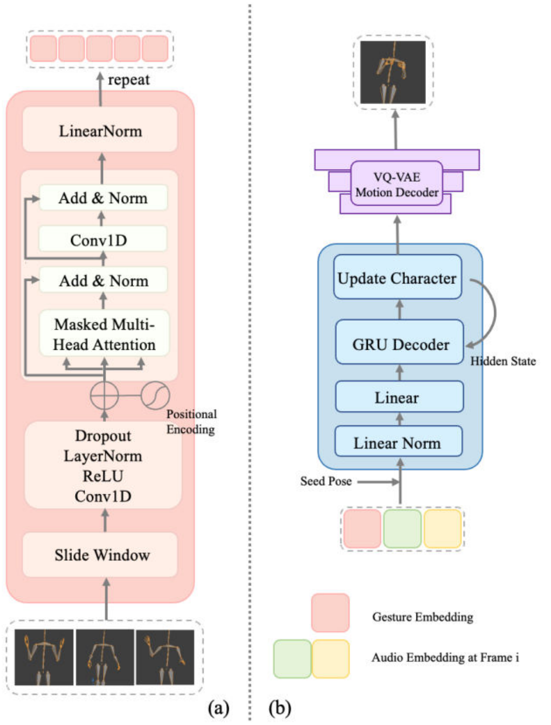

Fig.2: Architectures of the Gesture Encoder and the Gesture Generator.

As shown in Fig.2(a), the obtained feature sequence is input into a 1D convolutional layer, an activation function, and a normalization layer to obtain the feature embedding vector. This vector is then connected to a multi-head self-attention layer, with residual connections and normalization applied after each layer.

Gesture Generator. As illustrated in Fig.2(b), the gesture generator is a conditional autoregressive model comprising two layers of GRU. This model integrates audio embedding, gesture embedding, and the previous frames gesture prediction to predict the gestures for the new frame. The output of this circular decoder is the translation and rotation of joints with their velocity, as well as the translation and rotation of the root. Consequently, the final output sequence is $pred = [\rho_p, \rho_r, v_p, v_r, o_p, o_r, \epsilon_p, \epsilon_r]$, where ρ_p, ρ_r, v_p, v_r represent the translation and rotation of joints and their velocity, ϵ_p, ϵ_r represent the translational and rotational velocity of the root, and o_p, o_r denote the position and orientation of the root, which helps stabilize the output position of the model.

When calculating the gesture for each frame, the GRU output is initially denormalized. The predicted root translation and rotation velocity are utilized for root transformation, which is then combined with the previously generated pose. After normalization, the GRU is then input, followed by accessing the trained VQ decoder to enhance the semantic relevance of the generated data.

Loss Function. The loss function consists of two components: the reconstruction loss for pose generation and the Kullback-Leibler (KL) divergence loss between the predicted distribution of the gesture encoder q(z |e) and the multivariate Gaussian distribution p(z) before prediction. The overall loss is:

$$L = L_{rec} + D_{KL}(q(\boldsymbol{z}\,|\boldsymbol{e})||p(\boldsymbol{z})) \quad (4)$$

The reconstruction loss consists of the following components:

$$L_{rec} = \lambda_p L_p + \lambda_r L_r + \lambda_{vp} L_{vp} + \lambda_{vr} L_{vr} + L_{dp} + \lambda_{dr} L_{dr} \quad (5)$$

The first four items are the mean square error loss (MSE) of joint position, rotation, translational velocity, and rotational velocity between the ground truth and the predict, and the last two items are the mean error loss (MAE) of acceleration of the rotation between the ground truth and the predict.

IV. EXPERIMENTS

Datasets. In this paper, we evaluate the proposed method using the largest high-quality speech-gesture dataset BEAT. BEAT was constructed using a commercial motion capture system with 16 cameras recording the conversation and self-talk processes. Gestures are categorized into four types, and emotions are divided into seven categories. The dataset is primarily in English but also includes four other languages, with 30 speakers from ten countries contributing to the dataset. This includes facial expressions and body movements. The dataset contains approximately 76 hours of motion capture and speech on various topics.

Data preparation. The audio file is 16000hz. We use a 20 ms window size to extract speech features, thus generating 30 fps of data. We down-sample the motion capture data from 60 fps to 30 fps to match the speech features. We normalize all speech and joint positions by mean and variance. For transcript text, we extract them through the open-source ASR model. We trained all models at 30 fps.

Baselines. Our method is compared with the following methods: ZeroEGGS and CaMN. They are two representative approaches for co-speech gesture generation. CaMN integrates multimodal information such as audio, text, facial expression, and speaker ID in a cascade structure, leading to a more comprehensive generation effect. ZeroEGGS focuses more on the processing of generating actions and rhythm adaptation. DiffuseStyleGesture, based on the diffusion model, achieves high-quality generation.

Training. The whole training process is divided into two steps. First, we use the Adam optimizer to train the VQ-VAE model with 400 epochs. In the second step, we use the Adam optimizer to generate 160000 epochs of the model in the batch size 256, with a learning rate of 1e-5 and a learning rate decay of 0.995.

Users study. For generative tasks, given the absence of definitive criteria, human subjective evaluation is the most critical method of assessment. Consequently, human subjective assessment is the primary method used to evaluate our approach. We recruited 20 volunteers who were tasked with grading the slices based on the following four criteria: (1) naturalness, (2) rhythm, (3) diversity, and (4) semantics. Each video is cut into 15-20s clips. Each criterion was scored on a scale of 1 to 10, with 1 being the lowest and 10 being the highest rating, indicating the worst to the best performance, respectively. As shown in Table.I, our model achieves the best scores in semantics, naturalness and rhythm.

Table.I: Users study on BEAT. Our CASG performs best in the term of semantics, naturalness and stability.

	Semantics	Naturalness	Diversity	Rhythm	Average
ZeroEGGS	7.28	8.58	7.86	9.12	8.21
CaMN	8.40	8.26	7.48	7.39	7.88
DiffuseStyleGesture	7.42	8.84	8.65	9.50	8.63
Ours	**8.62**	**8.91**	8.06	**9.63**	**8.77**

Quantitative Evaluation. FGD [16] and BeatAlign [17] are utilized as quantitative evaluation metrics. FGD is an indicator that quantifies the discrepancy between the generated pose and a reference pose. We have developed a specialized network for FGD, utilizing a pre-trained LSTM-based autoencoder to extract features that can capture the dynamic changes inherent in time series data. Additionally, BeatAlign is employed to assess the synchronization between the generated gesture and the audio beat. This synchronization is evaluated by calculating the temporal alignment between the generated gesture and the audio beat. As shown in Table.II, our model still achieves the best performance.

Table.II: Quantitative Evaluation on BEAT.

	FGD↓	BeatAlign↑
ZeroEGGS	80.1	0.849
CaMN	123.7	0.783
DiffuseStyleGesture	79.4	0.941
Ours	**75.9**	**0.962**

V. Conclusion

In this article, we introduce CASG, a method designed to enhance the semantic expression capabilities of co-speech gesture generation. We propose a semantic enhancement module (SEM) that can fast generate gestures with both semantic and rhythmic qualities. The refined loss function ensures the stability of the generated poses. Experiments demonstrate that our model achieves competitive results in terms of semantics, diversity, and stability.

Large Language Model (LLM) have emerged in the field of natural language processing in recent years, which exhibits strong deep language understanding and generalization abilities, significantly improving the performance of various downstream tasks. In future work, we plan to leverage LLMs' powerful semantic comprehension to further enhance the effectiveness of co-speech gesture generation.

References

[1] Liu H, Zhu Z, Iwamoto N, et al. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 612-630.

[2] Ghorbani S, Ferstl Y, Holden D, et al. ZeroEGGS: Zeroü∟shot Exampleü∟based Gesture Generation from Speech[C]//Computer Graphics Forum. 2023, 42(1): 206-216.

[3] Yang S, Wu Z, Li M, et al. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models[J]. arXiv preprint arXiv:2305.04919, 2023.

[4] Zhi Y, Cun X, Chen X, et al. Livelyspeaker: Towards semantic-aware co-speech gesture generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 20807-20817.

[5] Ao T, Zhang Z, Liu L. Gesturediffuclip: Gesture diffusion model with clip latents[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-18.

[6] Tevet G, Gordon B, Hertz A, et al. Motionclip: Exposing human motion generation to clip space[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 358-374.

[7] Gao N, Zhao Z, Zeng Z, et al. GesGPT: Speech Gesture Synthesis With Text Parsing from ChatGPT[J]. IEEE Robotics and Automation Letters, 2024.

[8] Yang S, Xu Z, Xue H, et al. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 7945-7949.

[9] Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.

[10] Zhang J, Zhang Y, Cun X, et al. Generating human motion from textual descriptions with discrete representations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14730-14740.

[11] Jiang B, Chen X, Liu W, et al. Motiongpt: Human motion as a foreign language[J]. Advances in Neural Information Processing Systems, 2024, 36.

[12] Gong K, Lian D, Chang H, et al. Tm2d: Bimodality driven 3d dance generation via music-text integration[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9942-9952.

[13] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[14] Chen S, Wang C, Chen Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[15] Zhang H, Starke S, Komura T, et al. Mode-adaptive neural networks for quadruped motion control[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 1-11.

[16] Yoon Y, Cha B, Lee J H, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity[J]. ACM Transactions on Graphics (TOG), 2020, 39(6): 1-16.

[17] Li R, Yang S, Ross D A, et al. Ai choreographer: Music conditioned 3d dance generation with aist++[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13401-13412